DOCUMENT RESUME

ED 040 881 SE 009 056

AUTHOR Hater, Mary Ann; Kane, Robert B.
TITLE The Cloze Procedure as a Measure of the Reading Comprehensibility and Difficulty of Mathematical English.
INSTITUTION Purdue Univ., Lafayette, Ind.
PUB DATE [70]
NOTE 25p.

EDRS PRICE EDRS Price MF-$0.25 HC-$1.35
DESCRIPTORS *Cloze Procedure, Mathematical Vocabulary, *Mathematics Education, *Reading Comprehension, *Reading Difficulty, Research, Secondary School Mathematics

ABSTRACT

The purpose of this study was to validate the cloze procedure as a measure of the comprehensibility and difficulty of mathematical English. The authors point out that the cloze technique cannot readily be applied to mathematical English as it can to ordinary English since this technique is not defined to include deletions of mathematical symbols, and mathematical English has no definite ordering of words. Results supported the hypothesis that cloze tests over mathematical English passages are highly reliable measures and valid predictors of the reading comprehensibility of mathematical English passages for grades 7-12. There was also sufficient evidence to suggest the conclusion that cloze tests are valid predictors of reading difficulty for mathematical English passages at these grade levels. (Author/FL)

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

ED040881

The Cloze Procedure as a Measure of the Reading Comprehensibility and Difficulty of Mathematical English

Sister Mary Ann Hater
Robert B. Kane
Purdue University

The cloze technique has been used to measure constructs in both written and oral communication for English and, in a few cases, for other languages. Written cloze tests, the only cloze tests discussed in this study, are constructed by deleting certain words or symbols from passages and replacing them with blanks. The subject attempts to complete the passages. His score for each passage is the number of responses which match the deleted material.

Results from research studies (Taylor, 1953, 1954, 1957; Bormuth, 1962, 1963, 1966; Fletcher, 1959; Jenkinson, 1957; Rankin, 1957; Friedman, 1964; Gallant, 1965) indicate that cloze tests are reliable and valid measures of reading comprehensibility and difficulty for ordinary English passages.

Kane (1967, 1969) pointed out differences between ordinary English (OE) and mathematical English (ME). Although the cloze technique has been validated as a measure of comprehensibility and difficulty of OE passages, it has not been validated for use with ME. Thus, the cloze technique cannot be used indiscriminately as a measure of reading comprehensibility and difficulty for ME.

The purpose of this study is to validate the cloze technique as a measure of the comprehensibility and difficulty of ME.

The cloze technique cannot readily be applied to ME since (1) the cloze technique is not defined to include deletions of mathematical symbols, and (2) the cloze technique requires an ordering of words, but unlike OE which has a one dimensional ordering of words on a page, ME has a two dimensional arrangement for which no definite ordering is defined.

Hater (1969) defined word-token and math-token and established an ordering for tokens which follows the ordering of verbal expressions of mathematical symbols. Word-tokens are graphic representations appearing in ME which have phoneme-grapheme relationships with spoken words. In general, a word-token is a unit of alphabetic signs having lexical meaning and separated from surrounding context by spaces, (Examples: circle, the, following). In general, math-tokens are symbols unique to mathematics which do not have phoneme-grapheme relationships with spoken words, (Examples: !, ∧, +). For this study, the cloze technique was adapted to be used with mathematical English passages by using these definitions of tokens and ordering.*

SE 009 056

Hypotheses

Hypotheses tested in this experiment were divided into two sections (1) major hypotheses, and (2) hypotheses which related to the design of the experiment.

*For details concerning these definitions see Hater (1969).

Major Hypotheses:

1. A cloze test over a ME passage is a reliable measure. The comprehension tests used in this study are reliable criterion measures.
2. A cloze test (Forml) over a ME passage is a valid predictor of reading comprehensibility.
3. A ranking by means on cloze tests over a ME passage is a reliable measure. The ranking by means on the comprehension tests used in this study is a reliable criterion measure.
4. A cloze test over a ME passage is a valid predictor of reading difficulty.

Related Hypotheses:

1. There is no difference between the means of the different forms of cloze tests over the same ME passage.
2. The cloze test treatments do not sensitize subjects and affect responses on comprehension tests.
3. Results on comprehension tests are affected by the reading of the passages before responding to the comprehension tests.

Design and Procedures

This study was designed to test each hypothesis by at least one verification and where possible by cross-validation. The basic materials consisted of five passages from mathematics books. Cloze and comprehension tests were written over these passages and then administered to subjects in grades 7 through 10.

Passages. Five ME passages (P(1) through P(5)) were chosen. P(1), a unit on Matrices, employed a discovery approach. P(2) through P(5), units on the Metric System, Matrices, Statistics, and Logic, used mixtures of definitional and explanatory material. Passages were lengthened or shortened to approximately 700 tokens. Exercises were eliminated but most of the questions which appeared in context were included. Pictures and graphs were included.

Cloze Tests. Five cloze tests were constructed for each of the five ME passages. Form i (i = 1, 2, 3, 4, 5) began with the deletion of the ith token. Every fifth token after that was deleted until 130 tokens were deleted. For each ME passage, all tokens, except those deleted, appeared in each cloze form.

Deleted tokens were replaced by red blanks. Red was used to differentiate blanks used to represent tokens from vincula used in fractions. Subjects were to infer and to write responses to blanks by reading and studying the contextual clues of the passages. Two-sized blanks were used depending on whether adjacent tokens were word-tokens or math-tokens. Tokens which appeared in pictures, diagrams, and charts were eligible for deletion.

Criterion Measures. The comprehensibility of Passage x was defined to be greater for Subject 1 than for Subject 2 if and only if the test score on a comprehension test over Passage x was greater for Subject 1 than for Subject 2. Passage x was defined to be more difficult than Passage y if and only if the mean of the comprehension test results for Passage x was less than the mean of the

comprehension test results for Passage y. Although there are problems in the use of comprehension tests as criterion measures of comprehensibility and difficulty for ME, at this stage there is no instrument generally accepted to be better.

A 28-item multiple-choice test with five alternative responses to each item was written over each of the ME passages. Directions and test item stimuli contained (1) no math-tokens which were not included in the ME passages being tested or in elementary mathematics textbooks, and (2) no word-tokens which were not in the ME passages or in Column G from AA through 10 of The Teacher's Word Book of 30,000 Words (1944), which is a listing of frequency of occurrence of words in general reading material which should be part of the permanent vocabulary by grades five and six.

Pilot studies were conducted to improve comprehension tests as valid criterion measures of comprehensibility and difficulty.

Subjects. After incomplete data from 117 subjects were eliminated from the experiment, data from 1717 subjects enrolled in Grades 7 through 10 were used in the final analysis. Subjects were enrolled in five grade schools and three high schools in Cincinnati, Dayton, Springfield, and Lincoln Heights, Ohio.

Treatments and Assignment of Subjects to Treatments. The data collection took place on three days; on the first day a cloze test was completed by subjects, six days later a ME passage was studied, and on the next day the ME passage was returned to be reviewed for ten minutes after which a comprehension test was taken. Time allotments for these activities were 55 minutes, 40 minutes, and 45 minutes respectively.

For each of the five comprehension tests, there were five experimental groups. Each subject was randomly assigned to one of the twenty-five groups. Following is the design for the comprehension test over P(1). The designs for the other comprehension tests were the same.

	Day 1	Day 2	Day 3
Group 1:	Cloze Test P(1), Form (1)	Read P(1)	Comprehension Test P(1)
Group 2:	Cloze Test P(1), Form (4)	Read P(1)	Comprehension Test P(1)
Group 3:	Cloze Test P(1), Form (5)	Read P(1)	Comprehension Test P(1)
Group 4:	Cloze Test P(i),	Read P(1)	Comprehension Test P(1)
Group 5:	Cloze Test P(j),	Read P(j)	Comprehension Test P(1)

$i \neq j \neq 1$

Cloze tests over P(1), Form 2 and 3, were administered to subjects of other experimental groups, thus each of the five cloze forms over each passage was administered to one group.

Some hypotheses required that Groups 1 through 5 be used together. However, because of the unequal group sizes, a Subgroup 1* of 32 subjects was randomly selected from Group 1 for each passage to test these hypotheses. In order to cross-validate, half of the subjects who were assigned to Group 1 for each passage were randomly assigned to form Validation Group 1'. The remainder of the subjects were assigned to Cross-Validation Group1". Table 1 gives the assignment of subjects

to groups.

Table 1

Assignment of Subjects to Groups

	Passage 1	Passage 2	Passage 3	Passage 4	Passage 5
Group 1	211	225	209	225	220
Group 2	31	28	33	32	31
Group 3	32	32	29	31	32
Group 4	30	31	35	34	33
Group 5	28	30	35	31	29
Group 1*	28	31	27	30	30
Group 1'	111	111	109	115	106
Group 1"	100	114	100	110	114

Scoring and Rescoring. All comprehension tests and cloze tests were machine scored. In general, a cloze response was considered correct if it exactly matched the writer's original response.* Independent rescoring took place. There was an almost total agreement on all of the cloze tests between the initial scoring and rescoring.

Analyses Used

Prior to testing the hypotheses, the F_{max} Test (Winer, 1962) was used to check homogeneity of variances for both cloze test results and comprehension test results ($\alpha = 0.01$). Also, distributions of subjects' scores on cloze tests (Form 1) and comprehension tests were graphed. Graphs were inspected for distributions; no statistical techniques were employed.

Hypotheses Related to Design. To test Hypothesis 1 ($\alpha = 0.01$), the relationships between different cloze forms over the same WME passages were tested by using two-way analysis of variances for unequal n's. The dimensions were five cloze form groups and four grades. The model included the interaction of forms and grades.

To test Hypotheses 2 and 3, the relationships between different comprehension test groups over the same passage were tested first by using two-way analysis of variances for unequal n's. The dimensions were five comprehension test groups and four grades. The model included the interaction of comprehension test groups and grades. Secondly, if there was a significant main effect for comprehension test groups ($\alpha = 0.05$) and an insignificant interaction between comprehension test groups and grades ($\alpha = 0.05$, Scheffe's method (1959) was used to make tests among the means. A priori decisions to use certain contrasts were made.

*For details concerning scoring see Hater (1969)

To test Hypothesis 2 ($\alpha = 0.01$), two contrasts were employed for each passage:

Contrast 1 (1, 0, 0, -1, 0): The mean of Group 1* was compared with the mean of group 4,

Contrast 2 (1/3, 1/3, 1/3, -1, 0): The mean of the means of Groups 1*, 2, and 3 was compared with the mean of Group 4.

To test Hypothesis 3 ($\alpha = 0.25$), one contrast was employed.

Contrast 3 ($\frac{1}{4}$, $\frac{1}{4}$, $\frac{1}{4}$, $\frac{1}{4}$, -1): The mean of the means of Groups 1*, 2, 3, and 4 was compared with the mean of Group 5.

Major Hypotheses. To test Hypothesis 1, K-R 20 was used. Reliabilities were obtained for Group 1 over all passages for both cloze tests and comprehension tests and for Groups 1* through 5 over all passages for both cloze tests and comprehension tests.

To test Hypothesis 2, first, test results for Validation Group 1' for each of the five passages were graphed separately, where the set of scores on the cloze test was the independent variable and the set of scores on the comprehension test was the dependent variable. The graphs were studied for relationships between the variables. Then, linear regression equations were obtained for each passage. The model was cross-validated using test results from Group 1" for each passage.

To test Hypothesis 3, product-moment correlations were obtained between means of cloze test results from Group 1' and Group 1" for all passages. Product-moment correlations were also obtained between means of comprehension test results from Groups 1' and 1" for each passage.

To test Hypothesis 4, the Pearson r was used to obtain a measure of the relationship between the means of the scores on the cloze and comprehension tests. Correlations were obtained between the means of:

a) cloze test results Group 1' for all passages and comprehension test results for all passages,
b) cloze test results Group 1" for all passages and comprehension test results Group 1" for all passages,
c) the mean of the cloze test results for Groups 1*, 2, 3, and 4 for all passages and the mean of the comprehension test results Groups 1*, 2, 3, and 4 for all passages.

Results

There was no evidence to suggest rejection of the hypothesis of homogeneity of variances of cloze test results for groups responding to different deletion systems over the same passages and for groups responding to a single deletion system over different passages. In general, there was no evidence to suggest rejection of the hypothesis of homogeneity of variances of comprehension test results over the same passage for groups responding to different cloze tests and for groups responding to one cloze form over different passages.

Frequency distributions of subjects' scores on cloze tests were unimodal for four of the five passages. For Passage 3, the distribution was irregular toward the center, but tapered toward the ends. Frequency distributions on comprehension tests were bimodal for P(1), P(2), and P(3). The modes to the right were smaller than the ones to the left. Skewness to the left appeared for P(1); skewness to the right appeared for the other four passages.

Hypothesis 1 Related to Design: The analysis of variance for the main effect of forms over each passage is summarized in Table 2. The observed values were 0.17, 1.40, 1.51, 0.95, and 2.02. In testing the null hypothesis of equal means, the value required for significance at the 0.01 level was $F_{.99}(4, 120) = 3.48$. Thus, there was no evidence to support the rejection of the hypothesis of equal means on cloze tests over the same passage for groups using different cloze forms.

Table 2

Summary of Analysis of Variance of Cloze Test Results
(Form x Grade)

	P(1)		P(2)		P(3)		P(4)		P(5)	
Source	df	F	df	F	df	F	df	F	df	F
Form	4	0.17	4	1.40	4	1.51	4	0.95	4	2.02
Grade	3	5.12	3	12.82	3	7.44	3	6.79	3	8.06
F x G	12	1.14	12	1.14	12	1.72	12	0.79	12	1.12
Error	134		129		129		143		138	

$F_{.99}(4, 120) = 3.48$

$F_{.95}(3, 120) = 2.68$

$F_{.95}(12, 120) = 1.83$

Hypotheses 2 and 3 Related to Design: The analysis of variance for the main effect of comprehension test groups is summarized in Table 3. The observed F values were 11.68, 2.56, 9.48, 8.59, and 12.95. In testing the null hypothesis of equal means, the value required for significance at the 0.05 level was $F_{.95}(4, 120) = 2.45$. Thus, there was evidence to support the rejection of the hypothesis of equal means on comprehension tests over the same passage. The observed F values for the interaction effect of comprehension test groups and grades were 0.85, 0.85, 1.15, 0.67, and 1.12. The value required for significance at the 0.05 level was $F_{.95}(12, 120) = 1.83$. Thus, there was no evidence to support the rejection of the hypothesis of no interaction of means on comprehension tests over the same passage for groups and grades.

Table 3

Summary of Analysis of Variance of Comprehension Test Results (Group x Grade)

	P(1)		P(2)		P(3)		P(4)		P(5)	
Source	df	F	df	F	df	F	df	F	df	F
Group	4	11.68	4	2.56	4	9.48	4	8.59	4	12.95
Grade	3	10.31	3	17.25	3	11.13	3	9.58	3	13.46
G X G	12	0.85	12	0.85	12	1.15	12	0.67	12	1.12
Error	129		132		139		138		135	

$F_{.95}(4, 120) = 2.45$

$F_{.95}(3, 120) = 2.68$

$F_{.95}(12, 120) = 1.83$

Since the assumptions were met, contrasts were obtained. The differences among comprehension test groups using contrast coefficient are summarized in Table 4. For Contrast 1, the observed values were 0.00, 0.00, 4.64, 0.42, and 0.05. The values required for significance at the 0.01 level were found by using $(q-1)F_{.99}(Q-1, d)$ where q was the number of comprehension test groups and d was the degrees of freedom for error. The critical value was 13.92 for 120 degrees of freedom. Thus, there was no evidence to support the rejection of the hypothesis of equal means on comprehension tests for subjects in different cloze groups.

For Contrast 2, the observed values were 0.21, 0.21, 3.93, 0.40, and 0.00. The critical value was 13.92 for 120 degrees of freedom for error. Thus, there was no evidence to support the rejection of the hypothesis of equal means on comprehension tests for subjects in different cloze groups.

For Contrast 3, the observed values were 45.81, 7.71, 29.45, 34.04, and 51.16. The critical value required for significance at the 0.25 level was 5.48 for 120 degrees of freedom for error. Thus, there was evidence to support the rejection of the hypothesis of equal means on comprehension tests for groups studying a passage before taking a comprehension test on it, and groups not studying the passage before taking a test on it.

Table 4

Summary of Differences Among Comprehension Test Groups
Using Contrast Coefficients

Passage	d^a	$(Psi/Sigma)^2$		
		1, 0, 0, -1, 0 Contrast 1	1/3, 1/3, 1/3, -1, 0 Contrast 2	$\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, -1$ Contrast 3
1	129	0.00	0.21	45.81
2	132	0.00	0.21	7.71
3	139	4.64	3.93	29.45
4	138	0.42	0.40	34.04
5	135	0.05	0.00	51.16

[a] d = Degrees of freedom for error
$(q-1)F_{.99}(q-1, 120 = 13.92$ where $q-1 = 4$
$(q-1)F_{.75}(q-1, 120) = 5.48$ where $q-1 = 4$

Main Hypothesis 1: Table 5 gives the means, standard deviations, and reliabilities of cloze tests, Form 1. The reliabilities of cloze tests (Form 1) for each passage were 0.94 for P(1), 0.93 for P(2), 0.96 for P(3), 0.95 for P(4), and 0.96 for P(5). Table 6 gives the means and reliabilities of cloze tests for the five forms over each passage. The range of reliabilities for the five cloze forms were 0.94 to 0.97 for P(1), 0.91 to 0.96 for P(2), 0.95 to 0.97 for P(3), 0.94 to 0.97 for P(4), and 0.96 for P(5).

Table 5

Means, Standard Deviations, and Reliabilities
of Cloze Tests for Form 1

Passage	Unweighted Mean	Standard Deviation	K-R 20
1	69.97	19.57	0.94
2	56.11	16.35	0.93
3	72.01	23.69	0.96
4	49.60	20.29	0.95
5	62.98	21.99	0.96

Table 6

Means and Reliabilities of Cloze Tests for Five Forms

	Passage									
	1		2		3		4		5	
Form	Mean	r	Mean	r	Mean	r	Mean	r	Mean	r
1*	73.00	0.95	51.94	0.91	25.12	0.97	18.61	0.95	22.95	0.96
2	76.15	0.94	44.77	0.96	61.19	0.95	51.86	0.97	53.47	0.96
3	76.27	0.95	51.89	0.94	65.86	0.97	42.77	0.96	59.45	0.96
4	73.87	0.97	51.25	0.92	67.03	0.95	51.77	0.96	64.74	0.96
5	73.03	0.96	46.19	0.95	61.00	0.95	47.13	0.94	67.03	0.96

Table 7 gives the means, standard deviations, and reliabilities of comprehension tests for Group 1 of each passage. The reliabilities of comprehension tests were 0.89 for P(1), 0.82 for P(2), 0.88 for P(3), 0.81 for P(4), and 0.86 for P(5). Table 8 gives the means and reliabilities for the five comprehension test groups over each passage. The ranges of reliabilities, excluding group 5 for each passage, were 0.89 to 0.92 for P(1), 0.78 to 0.87 for P(2), 0.84 to 0.88 for P(3), 0.79 to 0.86 for P(4), and 0.84 to 0.88 for P(5). The reliabilities for Group 5 were 0.82 for P(1), 0.79 for P(2), -0.08 for P(3), 0.22 for P(4), and 0.21 for P(5).

Table 7

Means, Standard Deviations, and Reliabilities of Comprehension Tests for Group 1

Passage	Unweighted Mean	Standard Deviation	K-R 20
1	16.99	6.21	0.89
2	12.45	5.31	0.82
3	11.86	6.01	0.88
4	11.15	5.11	0.81
5	14.70	6.06	0.86

Table 8

Means and Reliabilities of Comprehension Tests for Five Groups

	Passage									
	1		2		3		4		5	
Group	Mean	r	Mean	r	Mean	r	Mean	r	Mean	r
1*	16.57	0.92	12.74	0.78	11.11	0.88	11.00	0.84	14.03	0.88
2	17.23	0.91	12.18	0.83	13.21	0.87	10.72	0.86	14.03	0.87
3	17.56	0.92	13.81	0.87	11.93	0.84	11.10	0.84	14.97	0.84
4	16.73	0.89	12.90	0.81	13.89	0.88	10.41	0.79	14.52	0.88
5	8.86	0.82[a]	10.17	0.79[a]	7.54	-0.08[a]	5.48	0.22[a]	7.03	0.21[a]

[a]Subjects did not read passage before taking comprehension test over it.

Main Hypothesis 2: From the graphs of the data, the model $y = ax^2 + bx + c$ was accepted to represent the data, where x denotes results on the cloze test and y denotes results on the comprehension test over a specific passage. The analysis of data using this model is summarized in Table 9. The observed values of F for significance of regression were 35.63, 61.72, 41.67, 57.69, and 82.82. The value required for significance at the 0.001 level was $F_{.999}(2, 120) = 7.31$. Thus, there was evidence to support the hypothesis of significance of regression.

For all five passages the correlation of x^2 with y was greater than the correlation of x with y. Thus, x^2 was entered as the first independent variable in each equation. Partial $\underline{t}$s for x were 0.08, -0.74, -0.88, -0.62, and 1.31. In testing the null hypothesis (b = 0), the value of $\underline{t}$ required for significance at the 0.05 level was $\underline{t}_{.975}(100) = 1.98$. Since no observed $\underline{t}$ values exceeded the critical value there was no evidence to support the rejection of the hypothesis that b = 0. For each passage, bx was removed from the model. New equations were calculated using the model $y = ax^2 + c$ for the validation groups (See Table 10).

Table 9

Analyses of Data Using Multiple Linear Regression Model
$y = ax^2 + bx + c$ Over Each Passage for Validation Groups

	Passage				
	1	2	3	4	5
Regression Coefficient a	0.0002	0.0027	0.0020	0.0022	0.0009
Regression Coefficient b	0.0110	-0.0826	-0.1017	-0.0524	0.1065
Intercept	9.9436	7.9919	8.5639	7.4078	4.3883
R: Multiple Correlation Coefficient	0.63	0.73	0.66	0.71	0.79
R^2: Coefficient of Determination	0.40	0.53	0.44	0.51	0.62
Std. Error of Estimate	4.63	3.68	4.58	3.98	4.02
F for Sig. of Regression	35.63	61.72	41.67	57.69	82.82
df Error	108	108	106	112	103
Partial t for x^2	1.27	2.83	2.38	2.92	1.37
Partial t for x	0.08	-0.74	-0.88	-0.62	1.31

$F_{.999}(2, 60) = 7.76$

$F_{.999}(2, 120) = 7.31$

$t_{.975}(100) = 1.98$

$t_{.975}(200) = 1.97$

Table 10

Analyses of Data Using Linear Regression Model
$y = ax^2 + c$ over Each Passage for Validation Groups

	Passage				
	1	2	3	4	5
Regression Coefficient a	0.0013	0.0020	0.0013	0.0018	0.0017
Intercept	10.2927	5.7520	5.3171	6.1258	7.4688
R: Correlation Coefficient	0.63	0.73	0.66	0.71	0.78
R^2: Coefficient of Determination	0.40	0.53	0.44	0.51	0.61
Std. Error of Estimate	4.61	3.67	4.57	3.97	4.03
F for Sig. of Regression	71.90	123.40	82.76	115.62	162.82
df Error	109	109	107	113	104

$F_{.999}(1, 60) = 11.97$

$F_{.999}(1, 120) = 11.38$

For each passage, a second sample from the population was used. Correlation coefficients obtained between x^2 and y were higher than the correlations between x and y for these groups. F values were significant. The analysis of data using the model $y = ax^2 + c$ over each passage for the cross-validation groups is summarized in Table 11.

Table 11

Analyses of Data Using Linear Regression Model
$y = ax^2 + c$ Over Each Passage for Cross-Validation Groups

	Passage				
	1	2	3	4	5
Regression Coefficient a	0.0017	0.0023	0.0013	0.0016	0.0014
Intercept	7.9980	4.4859	3.7856	6.5560	8.1324
r = Correlation Coefficient	0.64	0.78	0.77	0.69	0.70
r^2 = Coefficient of Determination	0.41	0.60	0.59	0.48	0.49
Std. Error of Estimate	5.04	3.35	3.83	3.26	4.07
F for Sig. of Regression	68.86	170.47	140.29	100.17	109.13
df Error	98	112	98	108	112

$F_{.999}(1, 60) = 11.97$

$F_{.999}(1, 120) = 11.38$

After the regression equation for test results over each passage was obtained for both validation and cross-validation groups, the regression coefficients from each equation for one sample were applied as weights to test results from the other sample. The regression equation was calculated using weighted scores. There was evidence to support the hypothesis of significance of regression in each case, (see Table 12).

Table 12

Correlations and F Ratios Obtained by Double Cross-Validation Using Model $y = ax^2 + c$

Group	Multiple Correlation Coefficients				
	1	2	3	4	5
Sample 1 Weights 1	0.63	0.73	0.66	0.71	0.78
Sample 2 Weights 1	0.63	0.76	0.81	0.69	0.64
Sample 2 Weights 2	0.64	0.78	0.77	0.69	0.70
Sample 1 Weight 2	0.60	0.70	0.66	0.69	0.71
	F for Significance of Regression				
Sample 2 Weights 1	63.00	154.82	180.40	96.89	78.96
df Error	98	112	98	108	112
Sample 1 Weights 2	62.01	107.45	84.36	105.55	107.35
df Error	109	109	107	113	104

Note: Sample 1: validation group; Sample 2: cross-validation group; Weights i refers to weights from linear regression equation using Sample i.

$F_{.999}(1, 60) = 11.97$

$F_{.999}(1, 120) = 11.38$

Although F values were lower using the model $y = bx + c$ than for the model $y = ax^2 + c$, they were highly significant. The correlations obtained using the model $y = bx + c$ are summarized in Table 13. For each passage, correlations obtained from the different sample groups were compared. Since the t statistic cannot be used with correlations ($r \neq 0$), z was used (Ostle, 1963). Obtained z values were 0.17, 0.61, 0.97, 0.53, and 1.27. In testing the null hypothesis of equal correlations, the value of z required for significance at the 0.05 level was 1.96. Thus, there was no evidence to suggest rejection of the hypothesis of equal correlations.

Using a z transformation, the average correlation of the correlations obtained using the validation groups for each passage was 0.69. Using a z transformation, the upper and lower confidence limits ($\alpha = 0.05$) were found using the cross-validation groups. In each case, the average correlation for the validation group was contained within the confidence interval using the cross-validation group.

Table 13

Correlations Obtained Using Model $y = bx + c$

	Passage				
	1	2	3	4	5
r_1 (Validation Groups)	0.62	0.71	0.64	0.69	0.78
r_2 (Cross-validation Groups)	0.64	0.75	0.71	0.65	0.70
z values	0.17	0.61	0.97	0.53	1.27
Upper Confidence Limit[a]	0.74	0.82	0.80	0.74	0.79
Lower Confidence Limit[a]	0.50	0.65	0.61	0.52	0.60

[a] $\alpha = 0.05$ with cross-validation groups

$z_{.975} = 1.96$

Main Hypothesis 3: Groups 1' through 5' and 1" through 5" were the two independent samples from the population used to rank the cloze tests and comprehension tests. Table 14 includes the means and reliability of cloze test rankings across sample groups. The product-moment index of relationship between the two sets of means was 0.99. Table 15 includes the means and reliability of comprehension test rankings across sample groups. The product-moment index of relationship between the two sets of means was 1.00 (rounded from 0.9983).

Table 14

Means and Reliability of Cloze Test Rankings Across Sample Groups

Passage	Means[a]		Correlation
	Validation	Cross-Validation	r^b
1	69.48	70.34	0.99
2	56.50	55.89	
3	71.35	72.77	
4	52.02	47.05	
5	62.73	62.63	

[a]Means are weighted according to the number of subjects in each grade over each passage

[b]r = Product-moment coefficient

Table 15

Means and Reliability of Comprehension Test Rankings Across Sample Groups

Passage	Means[a]		Correlation
	Validation	Cross-Validation	r^c
1	17.13	16.75	1.00[b]
2	12.80	12.20	
3	12.38	11.32	
4	11.63	10.66	
5	14.87	14.33	

[a]Means are weighted according to the number of subjects in each grade over each passage.

[b]Rounded from 0.9983

[c]r = Product-moment coefficient

Main Hypothesis 4: Figure 1 contains a comparison of the means of cloze tests and comprehension tests for the validation groups, cross-validation groups, and Groups 1* through 5* for all passages. For all three sets of groups, four of the five passages were ranked the same by the cloze test means and comprehension test means. However, for the fifth passage (P(3)), the cloze test mean and comprehension test mean ranked the passage quite differently. The cloze test ranked the passage easier than the comprehension test for all three sets of groups.

Correlations between the means of cloze test and comprehension test results are summarized in Table 16. The correlations were 0.54 for the validation groups, 0.51 for the cross-validation groups, and 0.83 for the combined groups. In testing the null hypothesis of no correlation, the values required for significance at the 0.05 and 0.10 levels were 0.88 and 0.81 respectively. Since the observed values did not exceed the critical values ($\alpha = 0.05$), there was little evidence to support the hypothesis of significant correlations.

Table 16

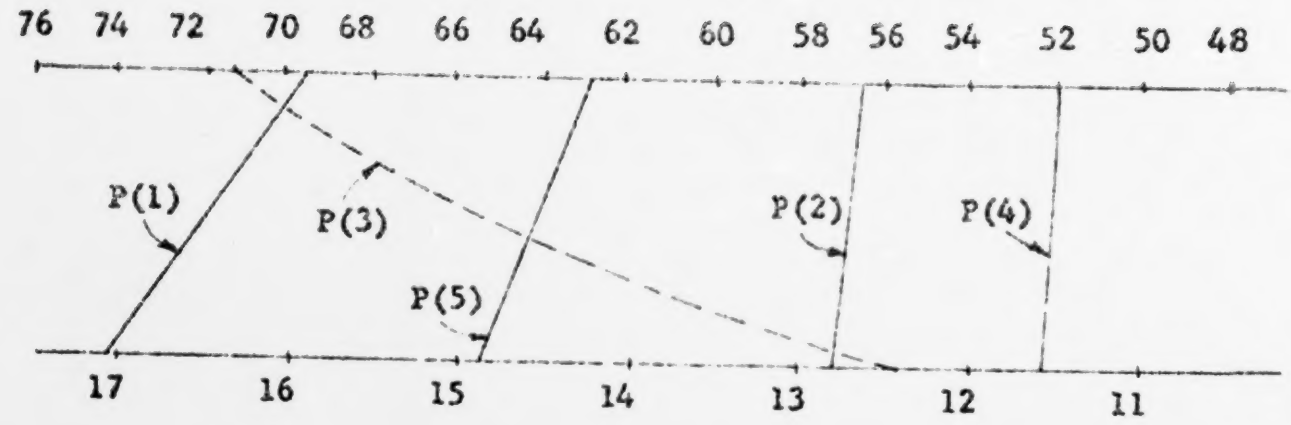
Correlations Between Ranking of Cloze Tests
and Ranking of Comprehension Tests

Groups	Correlation Coefficient r^a
Form 1: Validation Groups	0.54
Form 1: Cross-Validation Groups	0.51
Forms 1 to 5: Combined Groups	0.83

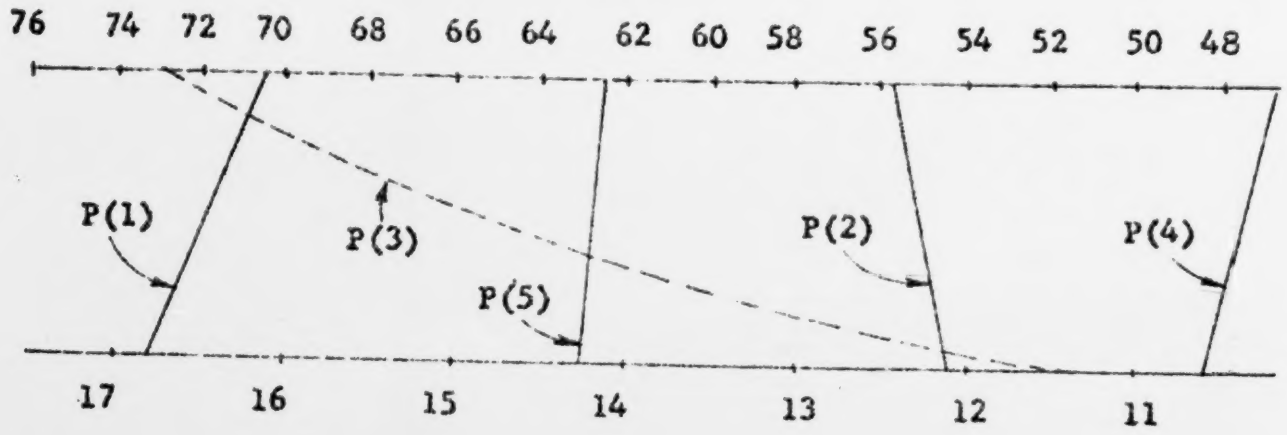
[a] r = Product-moment correlation

An r of .81 is significant at the 0.10 level

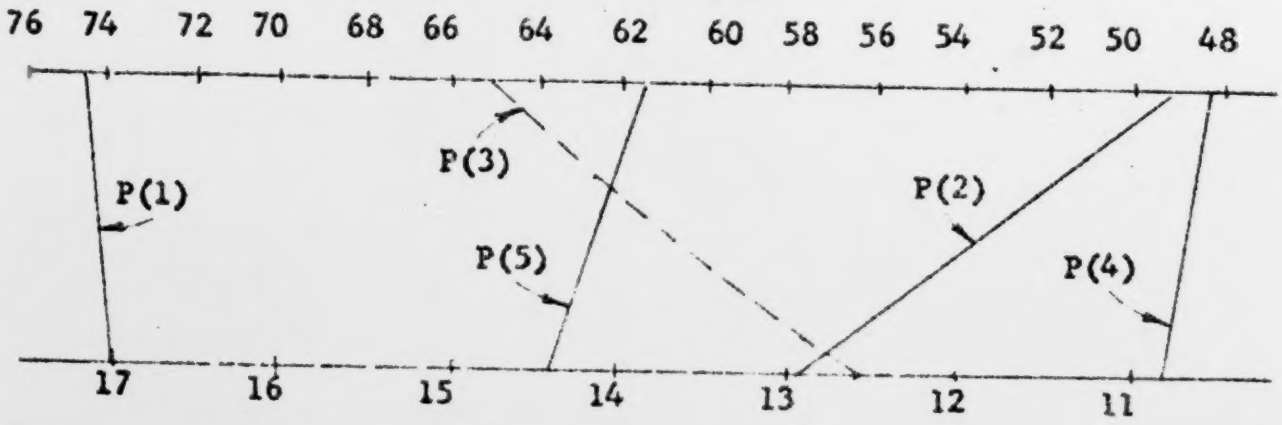
An r of .88 is significant at the 0.05 level



Validation Groups



Cross-Validation Groups



Five Groups

Note 1: P(i) refers to Passage i

Note 2: The top line in each set graphs cloze test means, the bottom line graphs comprehension test means.

FIGURE 1

COMPARISON OF MEANS OF CLOZE AND COMPREHENSION TESTS FOR DIFFERENT GROUPS

Conclusions

Conclusions Related to the Design. Three hypotheses were tested which checked factors related to the design of the experiment. First, the procedure of using only one cloze form to test the validity of cloze tests as measures of reading comprehensibility seemed justified. This conclusion resulted from analyses of all cloze forms over the same passage. There was no reason to believe that cloze forms over the same passages have unequal means and unequal variances.

A second factor which related to the design of the experiment was the effect of the cloze test treatment on comprehension test scores. There was no evidence to suggest differences in means of scores on comprehension tests for groups of subjects who responded to a cloze test over a passage different from the one tested by the comprehension test. Therefore, it was concluded that cloze test treatments did not unduly sensitize subjects and cause the subjects to respond differentially to comprehension tests as a result of completing cloze tests over the passages.

A third factor which was studied was the role of reading and studying passages on comprehension test results. It was hypothesized that, if the comprehension tests were tests of comprehension and not merely of former knowledge, there should be evidence of different responses for subjects who read and subjects who did not read the passages. Therefore, results for subjects who read and studied the passages over which comprehension tests were written were compared with results from subjects who did not read the passages.

For all five passages, the hypothesis of equal means on comprehension tests for these groups was rejected. For four of the five passages, the differences between means for groups who read and those who did not read were very large. Since the higher means were found for groups in which subjects read and studied the passages, it was concluded that the comprehension test scores were not only a result of background of subjects, but also a result of reading and studying the passages over which the tests were constructed. Thus, there was no evidence to reject the hypotheses related to the design of the experiments.

Results suggested that: (1) for each passage it is sufficient to use one form of cloze tests for validation of cloze tests as measures of reading comprehensibility, (2) the cloze test treatments do not sensitize subjects so as to affect responses on comprehension tests, and (3) reading and studying passages affects comprehension test results.

Conclusions Related to Main Hypothesis 1. In all instances reliability coefficients were greater than 0.90. Thus, it was concluded that cloze tests were highly reliable when reliability was measured by the K-R formula 20.

A comparison of reliabilities of responses by large and small groups was made over the same cloze tests. The maximum fluctuation of correlation coefficients was found for Passage 2, but the fluctuation was small. Thus, it was concluded that reliabilities obtained from cloze test results using small groups of approximately 30 subjects were similar to reliabilities obtained from cloze test results using larger groups of approximately 200 subjects.

A factor which undoubtedly contributed to the high reliabilities was the length of the tests. Only a few items had a difficulty level of 1.00 or 0.00. After eliminating the extreme ends of difficulty, each test included more than 100 items. With so many items in a range which allowed for discrimination by subjects, reliabilities were high.

Unlike multiple-choice items having n choices which allow correct responses to 1/n items to be due to guessing alone, correct guessing on cloze tests resulted from correct choices of words existing in the minds of the subjects or from tokens in close proximity to the missing tokens on the cloze tests, and not from correct guesses of alternative-responses on a test. Since the replacement set arising from both the subjects' background and the tokens on cloze tests was large, the error created by guessing a correct response was small. It appeared that the number of items replaced correctly by guessing alone was nearly zero for cloze tests. The fact that correct responses due to guessing were held at a minimum was one reason for the reliabilities of cloze tests being higher than the reliabilities of multiple-choice tests.

For four of five cloze tests, the average percent of subjects responding to the last five items on each test was similar to the average percent of subjects responding to a random selection of five items from the rest of the test. However, on Passage 4 the percents were different for the last five items and the random items respectively. The fact that some subjects quit before others could cause an increase in both the item discrimination values for the last items of the test and the test variances, thus resulting in an apparent increase in reliability.

In summary, some of the intrinsic characteristics of the cloze tests which seemed to contribute to high reliabilities were the length of the tests, the near-absence of the guessing-factor, and the distributions of item difficulty. One extrinsic factor which could have caused spurious reliabilities for the cloze test over one of the passages was the quitting-factor.

The K-R 20 coefficient for 28-item comprehension tests over all comprehension test groups in which subjects read the passages prior to taking the tests were greater than 0.77. It was concluded that comprehension tests were reliable criterion measures when reliability was measured by the K-R formula 20.

The coefficients for groups in which the subjects read passages different from those they were tested on ranged from 0.82 to -0.08. The low reliabilities were anticipated for these groups since errors due to guessing were magnified. The comprehension tests were written to test the amount of information gained through reading. The content for the passages was unfamiliar to the subjects in the experiment. Therefore, it was difficult for subjects to respond correctly to questions by using their former knowledge. Since questions were multiple-choice, correct responses to items could be a result of correct guesses. As a result, there was additional variation in responses from one item to another which lowered reliabilities.

The reliabilities of multiple-choice tests were high; however, they were not as high as the reliabilities of cloze tests. Two reasons which seemed to account for part of the differences in reliability coefficients were: (1) the length of the tests, and (2) the errors due to guessing which were present in the multiple-choice comprehension tests.

Conclusion Related to Main Hypothesis 2. The equations calculated for each passage using the model $y = ax^2 + bx + c$ showed that multiple correlations were large. However, for each equation, there was no evidence to suggest that the regression coefficients for the x terms were different from zero. Consequently, x was removed from the model. New equations were calculated using the model $y = ax^2 + c$.

As a method of further validation, a second sample from the population was tested on each passage. Using the new model, the multiple correlation coefficients and the amount of accountable variances were large and about the same sizes as with the original samples. Regression coefficients using one sample were applied as weights to the other sample. In all cases, multiple correlations were large.

Therefore, it was concluded that a model was found which described the relationship between cloze test and comprehension test results. Within the limitations imposed by comprehension tests, the model $y = ax^2 + c$ was appropriate for the data.

However, one of the limitations of this study was the distributions of the subjects' scores on comprehension tests. As was pointed out earlier, these distributions were skewed. The affect of the skewness on the relationship between cloze test and comprehension test scores is not to be discounted. Whenever two sets of scores are correlated, the size of the correlation can be restricted by different shaped distributions. Since the distributions of four comprehension tests were skewed toward the left in the scatter diagram depicting the relationship between comprehension test scores and cloze scores, y was skewed toward the lower end of the graph in the relationship between x and y. Consequently, a curvilinear relationship resulted which was not due necessarily to the variables being measured, but could have been due instead to the distributions of the criterion measures.

Another limitation, not independent of the first, which may have contributed to the curvilinear relationship was the guessing-factor. The possible range of subjects' scores on cloze tests was 131 with a negligible guessing-factor; the possible range of subjects' scores on comprehension tests was 29 with a high guessing-factor. Since there were 28 items of five responses each, the most likely minimum scores due to guessing on comprehension tests distributed around 5.6, not 0. As a subjects' knowledge of the content increased, fewer of his responses were guesses. As a result a floor was created at the lower left-hand corner of the scatter diagram.

This discussion of the effect of skewness and the guessing-factor on scatter diagrams of cloze test and comprehension test scores points out reasons which may have accounted for the curvilinear relationships between the two variables. A study of the graphs revealed that the linear and quadratic curves departed from one another in the lower left-hand region of the graphs, while in other regions the curves nearly coincided.

In conclusion, within the limitations imposed by the comprehension tests, model $y = ax^2 + c$ accounted for a little more variance than the model $y = bx + c$. However, if these limitations were eliminated, the simpler model $y = bx + c$ appeared

to be equally appropriate. The use of the comprehension tests with only 28 multiple-choice responses placed a restriction on the responses. Because of this limitation, it seemed appropriate not to reject the model $y = bx + c$ as the model of the prediction of comprehensibility from cloze test responses.

The amount of accountable variance using the model $y = bx + c$ was large. A further study of this model revealed that correlations for the validation and cross-validation groups over each passage were not significantly different. A correlation coefficient of 0.69 was found as the average correlation coefficient across all five passages for the validation groups. Confidence bands were placed on the correlation coefficients for the cross-validation groups. In each case the average correlation coefficient, 0.69, was contained within the confidence band. It was, therefore, concluded that the correlation coefficient of 0.69 was a good estimate of the relationship between the cloze tests and comprehension tests over passages.

Conclusions Related to Main Hypothesis 3. In general, it was found that rankings by means of cloze tests and comprehension tests are reliable measures since the rankings by means were nearly the same for different samples of the same population.

Conclusions Related to Main Hypothesis 4. Although the correlations between means of cloze tests and comprehension tests did not reach the desired level of significance, they were high (validation groups, 0.54; cross-validation groups, 0.51; and combined groups, 0.83). Because of their magnitude, it was concluded that a relationship between cloze tests and comprehension tests was probable, but that the strength of this relationship could not be determined because of the small sample of passages used in the study. Correlations were high enough to warrant further study of the relationship between cloze tests and comprehension tests.

An inspection of the means of cloze tests for the validation and cross-validation groups and the cloze tests over all five forms revealed that: (1) for Passage 1, the mean of Cloze Form 1 was smaller than the mean of the other forms over this passage, and (2) for Passage 3, the mean of Cloze Form 1 was larger than for the other forms over the passage. Therefore, when a single form was used, Passage 3 was ranked easier than Passage 1. However, when the five forms for each passage were averaged, Passage 1 was ranked easier than Passage 3.

In addition, using only one form resulted in a lower correlation with the criterion measure than using all five forms. In all cases, the mean of the comprehension test over Passage 1 was greater than the mean of the comprehension test over Passage 3. It was concluded that an average across the five cloze forms over a single passage was a better index of the passage difficulty and a better correlate with the criterion measure than the mean of only one cloze form.

For four of the five passages there were no inversions in the ranking of cloze tests and comprehension tests by means. However, for Passage 3 the mean of the cloze test scores was the largest or second largest mean of the five means, and the mean of the comprehension test scores was the second smallest mean. The topic of Passage 1 and Passage 3 was matrices. For both passages it appeared that the redundancy of matrix symbols and numbers in the passages made it easy to respond to the cloze tests.

In Passage 1, only one topic was presented, the multiplication of matrices. By contrast, in Passage 3, the topics presented included matrix addition, subtraction, multiplication, identity element, cancellation law, and other topics. Passage 1, 2, 4, and 5 were written to be used with school children, but Passage 3 was written as a summary for the general public. Passage 3 contained so many new topics that subjects were not able to learn and respond to a comprehension test over all of the topics at once without confusing the concepts. When subjects responded to topics one at a time on cloze tests, they did much better than when they sorted and responded to them on comprehension tests.

In conclusion, higher correlations were obtained when five cloze tests over a single passage were used than when only one cloze form was used. There were no inversions in the rankings of cloze tests and comprehension tests for four of the passages. Only for Passage 3 was there reason to suspect the relationship between cloze tests and comprehension tests.

Summary

Results of this study supported the hypothesis that cloze tests over mathematical English passages are highly reliable measures and valid predictors of the reading comprehensibility of mathematical English passages for the grades tested. An average linear correlation of 0.69 was found to represent the relationship between cloze test and comprehension test scores. This correlation may underestimate the relationship between the tests since a quadratic model accounts for more variance than a linear model. However, the additional variance may be due to guessing and subjects' distributions on comprehension tests.

Using double cross-validation techniques, it was found that correlations were similar for the different samples tested over the same passage and for samples tested over different passages. Therefore, conclusions concerning the use of cloze tests as predictors of comprehensibility were strengthened.

Cloze test means over mathematical English passages were ranked the same by different samples from the same population. Thus, it was concluded that the ranking of cloze tests is a reliable measure. Since the number of passages used in this study was only five, conclusions concerning the validity of cloze tests as measures of difficulty are tentative. Cloze test means were ranked the same as comprehension test means for four of the five passages. Consequently, there is enough evidence to suggest the probable conclusion that cloze tests are valid predictors of reading difficulty for mathematical English passages.

References

Bormuth, J. R. Cloze tests as a measure of comprehension ability and readability. (Doctoral dissertation, Indiana University) Bloomington, Ind.: University Microfilms, 1962. No. 62-2586.

Bormuth, J. R. Cloze as a measure of readability. International Reading Association Conference Proceedings, 1963, 8, 131-134.

Bormuth, J. R. Readability: A new approach. Reading Research Quarterly, 1966, 1, 79, 132.

Fletcher, J. E. A study of the relationships between ability to use context as an aid aid in reading and other verbal abilities. (Doctoral dissertation, University of Washington) Seattle: University Microfilms, 1959, No. 59-5483.

Friedman, M. The use of the cloze procedure for improving the reading comprehension of foreign students at the University of Florida. (Doctoral dissertation, University of Florida) Gainesville: University Microfilms, 1964, No. 64-11,533.

Gallant, R. Use of cloze tests as a measure of readability in the primary grades. International Reading Association Conference Proceedings, 1965, 10, 286-7.

Hater, M. A. The cloze procedure as a measure of the reading comprehensibility and difficulty of mathematical English. Unpublished Doctoral Dissertation, Purdue University, 1969.

Jenkinson, M. D. Selected processes and difficulties in reading comprehension. Unpublished Doctoral Dissertation, University of Chicago, 1957.

Kane, R. B. The readability of mathematical English. Journal of Research in Science Teaching, 1967-8, 5, 296-298.

Kane, R. B. The readability of mathematical textbooks: Revisited. Mathematics Teacher, 1969, in press.

Ostle, B. Statistics in research. (2nd ed.) Ames, Iowa: Iowa State University Press, 1963.

Rankin, E. F., Jr. An evaluation of the cloze procedure as a technique for measuring reading comprehension. (Doctoral dissertation, The University of Michigan) Ann Arbor: University Microfilms, 1957, No. 58-3722.

Scheffé, H. The analyses of variance. New York: John Wiley and Sons, 1959

Taylor, W. L. Cloze procedure: A new tool for measuring readability. Journalism Quarterly, 1953, 30, 414-438.

Taylor, W. L. Application of "cloze" and entropy measures to the study of contextual constraint in samples of continuous prose. (Doctoral dissertation, University of Illinois) Urbana: University Microfilms, 1954, No. 10-554.

Taylor, W. L. 'Cloze' readability scores as indices of individual differences in comprehension and attitude. Journal of Applied Psychology, 1957, 41, 19-26.

Winer, B. J. Statistical principles in experimental design. New York: McGraw-Hill, 1962.